ANALYSIS, INDEXING AND RETRIEVAL OF DIGITAL VIDEO: SOME NEW APPROACHES

Synopsis of the Ph.D. Thesis to be submitted

FOR THE AWARD OF THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

COMPUTER SCIENCE

By

SHARMILA KUMARI, M.

POST GRADUATE DEPARTMENT OF STUDIES AND RESEARCH IN COMPUTER SCIENCE

MANGALORE UNIVERSITY

MANGALAGANGOTHRI-574 199

KARNATAKA

INDIA

MARCH 2012

SYNOPSIS OF THE Ph.D. THESIS TO BE SUBMITED

TO

MANGALORE UNIVERSITY

FOR THE AWARD OF THE DEGREE OF DOCTOR OF PHILOSOPHY

IN

COMPUTER SCIENCE

Title: ANALYSIS, INDEXING AND RETRIEVAL OF DIGITAL VIDEO: SOME NEW APPROACHES

Name of the Candidate	SHARMILA KUMARI, M
Name of the Guide	Dr. B H SHEKAR
Date of Registration	21-09-2007
Registration No.	MU/EXB/Ph.D./CR.68/Comp.Sc./2007-08/E.13 Dated: 26.08.2008
Place of Work	POST GRADUATE DEPARTMENT OF STUDIES AND RESEARCH IN COMPUTER SCIENCE, MANGALORE UNIVERSITY, MANGALAGANGOTHRI-574 199, KARNATAKA, INDIA.

Abstract

Due to the unprecedented growth of digital media, there has been a growing demand for the design and development of effective and efficient algorithms for processing digital data. The digital data may be in the form of massive collection of images, video documents which are to be processed effectively to obtain the meaningful contents of digital data. In this context, plethora of algorithms have been developed by many researchers to analyze and retrieve the digital media contents. The indexing is used as an intermediate step in the retrieval process. The grand challenge faced by the researchers is to develop efficient storage/retrieval algorithms because of massive amount of digital data to be processed in most of the applications.

In this research work, we have developed some new algorithms to analyze and represent the more common database called *video database* which finds its applications in many of the computer vision based technology adopted fields such as visual surveillance, biometrics, crowd management, news video analysis, sports video analysis etc. The analysis, indexing and retrieval of desired video frame are the major phases addressed in this research work. We have devised some novel algorithms for effective analysis, efficient indexing and accurate retrieval of video data from a large collection of video databases. We have also developed methods for automatic segmentation of digital videos and demonstrate their suitability for indexing and retrieval.

The video database analysis is studied in two different perceptions in this work. In the first part of the research work, we have considered face based video analysis and understanding and hence devised efficient and accurate face representation schemes for effective video analysis. The local descriptors based algorithms namely PCA-SIFT, SIFT-FLD and Region Covariance Matrix descriptors are proposed which are found to be most accurate and useful for any video database processing. The higher dimensionality problem exists in the basic SIFT algorithm is eliminated through the identification of redundant information by employing principal component analysis in our PCA-SIFT and hence made efficient in terms of recognition time without

deterioration in recognition accuracy. On the similar line, in SIFT-FLD, a discriminative classifier model called linear discriminant analysis is employed on the gradient patches of image data for its minimal and discriminative representation and hence leads to compact representation of the data. Often, we found that the region covariance matrices are efficient too for any object based video data analysis, but fails to face based video analysis applications and hence we have devised the monogenic scale space based region covariance matrix descriptor which is found to be more accurate and efficient for face based video analysis and understanding.

In the second part of our work, we have addressed the video analysis and understating through the localization and interpretation of textual information. We have restricted our work only to localize the textual information as we felt that the subsequent works such as text recognition and indexing are well solved and an OCR engine could be used for recognizing the textual information. We have developed some new algorithms which utilize the gradient information of a video frame preceded by keypoints based analysis of image data for accurate localization of textual information. In order to overcome the problem of illumination, we have also developed some new techniques based on the phase of the discrete Fourier transform and discrete cosine transform. Experimental study has been conducted on the standard document databases to exhibit the performance of the proposed algorithms.

In addition, we have also addressed the problem of video segmentation by introducing our own video segmentation techniques namely color-SIFT and colour moments which are proved to be more accurate and efficient. The intermediate phase in video retrieval is the indexing mechanism which is useful for quick retrieval of the desired information. We have proposed a multi-dimensional and multi-level indexing structure called Universal B-Tree for storing the key frames which are identified during shot boundary detection and the representation mechanism chosen is based on the newly introduced local descriptor based face representation scheme.

All the proposed models are experimentally validated for their applicability on different domains by using standard face databases, viz. AT&T, YALE and UMIST face databases. We have also conducted experimentation on the standard ICDAR-

2003 database and TRECVID video database to exhibit the computational performance of the proposed algorithms and their suitability to video database analysis and understanding. Also, necessary theoretical backgrounds for all the proposed models have been provided. The superiority over several other models in terms of computational efficiency and recognition accuracy has been established through extensive experimental analysis.

1.0 INTRODUCTION

In this digital world, there has been a growing demand for video database processing in many computer vision applications. The video database is a storehouse of information on the various aspects related to the videos. The growing number of applications for video-intensive products, ranging from personal video recorders to multimedia collaborative systems, introduces many technical challenges behind the design of contemporary video database systems. Progress in video database analysis has been advanced by many research fields including information retrieval, image processing and computer vision, pattern recognition, multimedia database organization, multidimensional indexing, data mining, machine learning, and visualization, psychological modelling of user behaviour, man-machine interaction, among many others.

The analysis, indexing and retrieval of desired video frame are the major phases to be addressed in the design of any video database based applications. The plethora of algorithms have been developed by many researchers to analyze and retrieve the video contents. The indexing is used as an intermediate step in the retrieval process. The grand challenge faced by the researchers is to develop efficient storage/retrieval algorithms because of massive amount of video database to be processed in most of the applications. The complexity is further increased in the case of mini-computing devices such as mobile phones, iPods etc. However, the task of video analysis and retrieval is complicated by its nature because of the visual nature of video database that ends up with numerous interpretations of the same data. In this research work, we have devised some novel algorithms for effective analysis, efficient indexing and accurate retrieval of video data from a large collection of video databases. We have also studied methods for automatic segmentation of digital videos and demonstrate their suitability for indexing and retrieval.

2.0 VIDEO ANALYSIS, INDEXING AND RETRIEVAL: A REVIEW

With the fast growth of multimedia information, video database analysis, indexing and retrieval have attracted increasing attention in recent years. Many applications have emerged in areas such as video-on-demand, distributed multimedia systems, digital video libraries, distance education, entertainment, surveillance and geographical information systems (Manjunath et al., 2002). The video database analysis aims at obtaining a structured organization of the original video content and understanding its embedded semantics like humans do. The video indexing is the task of tagging meaningful video units obtained from video database analysis to enable convenient and efficient content retrieval. However, understanding the video contents is very difficult for a computer to emulate because of the limitations of machine perception under unconstrained environments and the unstructured nature of video database. Hence, devising an efficient video content analysis techniques is still an open research problem despite a large amount of effort in this area (Zhang et al., 1995; Yeung and Leo, 1997; Tsekeridou and Pitas, 2001).

So far, the principal approach to the problem of video analysis is to extract some low- to mid- level visual/textual features, then partially derive or understand the video semantics by analyzing and integrating these features. Fig. 1 shows a hierarchical video analysis structure where many popularly used features such as color, texture, shape, motion, shots (Lienhart, 1999), key-frames (Zhang, 1995), object trajectories, human faces (Tsekeridou and Pitas, 2001), as well as classified audio classes (Zhnag and Kuo, 1999) constitute the low- to mid-level indexing features. Obviously, a semantic gap still exists between the real video content and the video contexts derived from these features.

In our research work, we proposed to extract two types of video content features, namely, *video objects (faces)* and *video text*, preceded by shot detection techniques in the first stage. Several efficient representation schemes were introduced for semantic analysis of video database contents in the subsequent stages. The sports video database and news video database are considered to demonstrate the feasibility of the proposed algorithms, to facilitate the browsing based on facial information. In order to

exhibit the performance of the proposed text detection techniques, the news video database and document image database is considered to locate textual information present in a video frame for its content analysis.

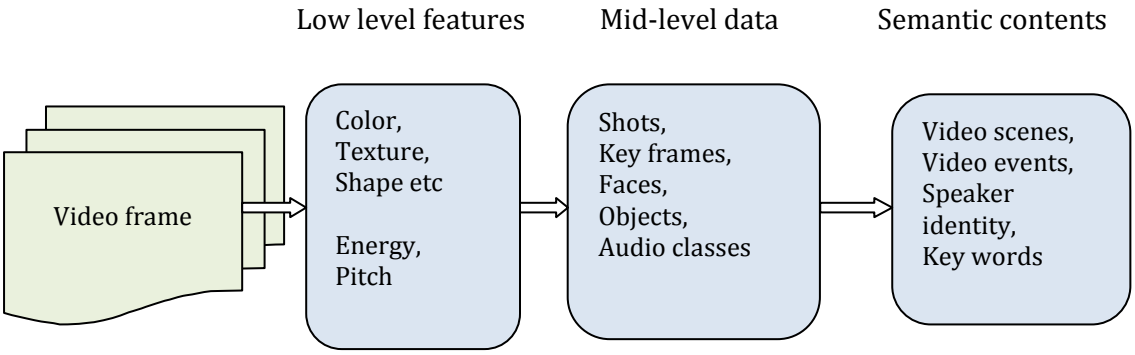

Fig.1 A Generic Video Analysis Structure

The rest of this pre-synopsis is organized as follows. We first present our shot detection algorithms which are used to segment the video database into collection of shots and key-frame is identified based on the pre-dominant visual features. In the next section, we present the face representation algorithms developed based upon appearance analysis and significant region analysis. The text detection models are given in section 4. The indexing mechanism developed based upon universal B-Tree is presented in section 5 followed by conclusion with avenues for future research work is given in section 6.

The experiments were carried out using the standard video databases such as TRECVID and its variants. The face representation algorithms are tested using standard face databases such as AT&T, UMIST and YALE. The text detection algorithms are tested using ICDAR-2003 document image database. All the reported algorithms are developed using MATLAB tool in WINDOWS environment using P-IV machine having 4 GB RAM.

3.0 PROPOSED MODELS

In this research work, we have addressed the problem of video analysis and indexing by introducing different algorithms for video segmentation, representation of segmented video and introduced an indexing mechanism for quick retrieval of desired frame from a large collection of video databases. The various stages in any video analysis and understanding are shown in Fig. 2. We have proposed algorithms for shot detection, key frame representation and indexing.

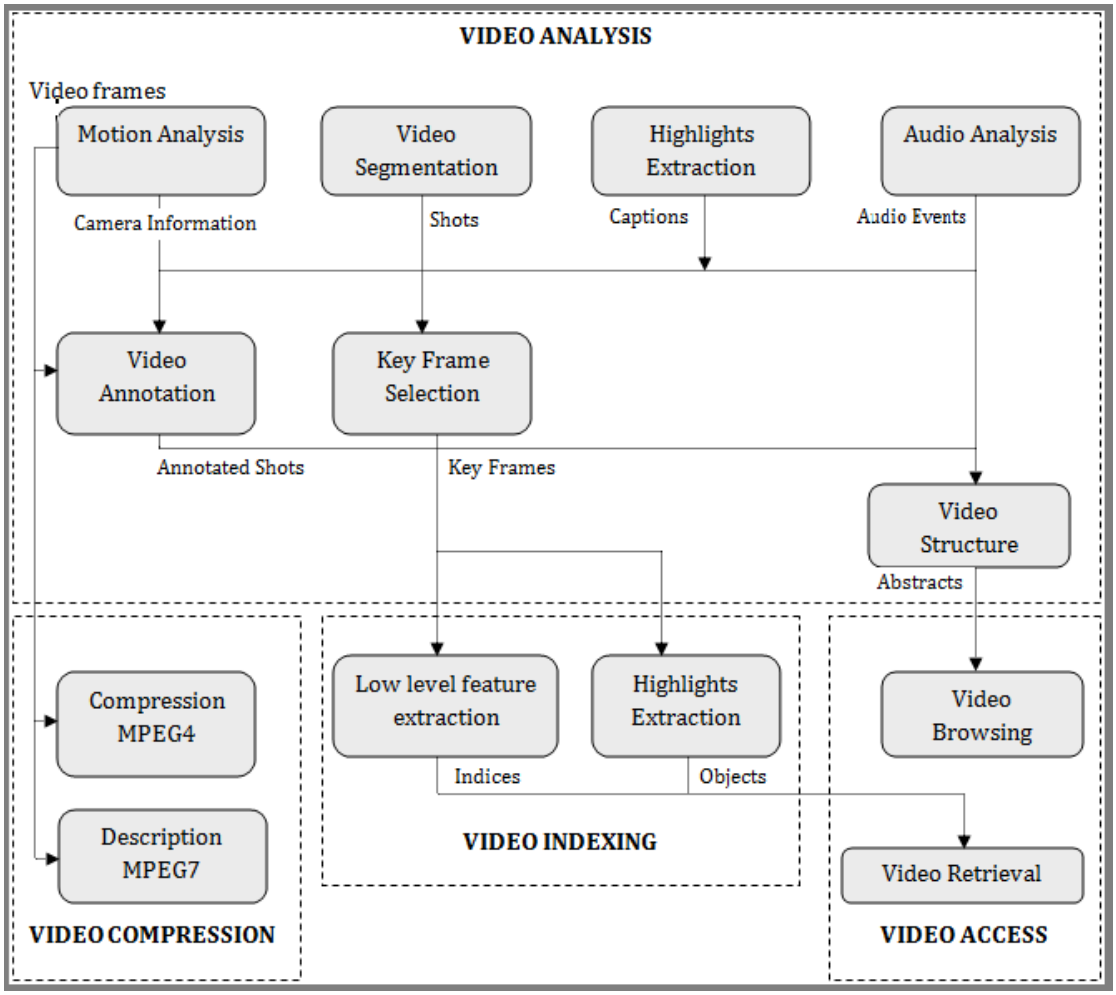

Fig. 2. Stages in video analysis and understanding

In the following sections, we briefly present each technique and highlight the merits of the proposed algorithm.

3.1 SHOT BOUNDARY DETECTION ALGORITHM[1]

The swift growth of video databases and the increasing demand to provide online access to these databases have brought content-based video retrieval (CBVR) to the attention of many researchers. Because manual indexing of archived videos is infeasible due to prohibitively high labor costs and massive amount of video databases, automation of video retrieval is essential to the online accessing of video databases. The massive amount of digital media coverage that results in huge video database generates difficulties in identifying and selecting desired information for semantic analysis of video content. Generally, video content can be represented by a hierarchical tree which contains shots, scenes, and events (Li and Kuo, 2003). The critical issue relies in identification of a continuously recorded sequence of image frames in such a huge data and hence, the continuous video bit-stream is segmented into a series of cascaded video shots, which are the basis for constructing high-level scenes and events with semantic meanings. Hence, shot detection (Rui et al., 1999), the identification of a continuously recorded sequence of image frames, is critical for semantic analysis of video content. Shot detection can generally be categorized into five classes: pixel-based, histogram-based, feature-based, statistics-based, and transform-based methods (Rui et al., 1999).

Quite a lot of approaches to shot boundary detection were proposed in the literature. The principle methodology of shot boundary detection is to extract one or more features from every 'n'th frame of a video sequence, to compute the difference of features for consecutive frames, and to compare these differences to a given threshold. Each time the threshold is exceeded, a shot boundary is detected.

---

[1] *An Efficient and Accurate Shot Boundary Detection Technique Based on Colour Moments,* ***International Conference on Advanced Computing & Communication Technologies****, pp. 77-80, Rohtak, India, Jan. 2011.*

Color-SIFT model: A robust and an accurate shot boundary detection algorithm, ***International Conference on Digital Image Processing (ICDIP-2010),*** *NEC, Nanyang Technological University, Singapore. Feb. 26-28, 2010.*

*An efficient and accurate shot boundary detection technique based on color moments****, International Journal of Artificial Intelligence and Knowledge Discovery****, Vol. 1(1), pp. 77-80, 2011.*

In our research work, we have devised shot detection algorithms suitable for accurate retrieval of a desired video frame from a huge collection of video databases. The proposed models are based upon (i) keypoint localization and (ii) color moments. The details of each model are given below in brief.

Color-SIFT based shot boundary detection technique: In this work, color-SIFT model is devised for shot boundary detection. Unlike scale invariant feature transform model that uses only grayscale information and misses important visual information regarding color, here we have adopted different color planes to extract keypoints which are subsequently used to detect shot boundaries. The basic SIFT model has four stages namely scale-space peak selection, keypoint localization, orientation assignment and keypoint descriptor and all these four stages were employed to extract key descriptors in each color plane. The proposed model works on three different color planes and a fusion has been made to take a decision on number of keypoint matches for shot boundary identification and hence is different from the color global scale invariant feature transform that works on quantized images. In addition, the proposed algorithm possess invariance to linear transformation and robust to occlusion and noisy environment. Experiments have been conducted on the standard TRECVID video database to reveal the performance of the proposed model.

Shot Boundary Detection Technique Based on Colour Moments: In this work, a shot boundary detection method based on primitives of colour moments is proposed. The YIQ colour model is used in this work. Here, we compute the colour moments of different order for each frame of the video. The feature vector is formed from the moments and compared between successive frames. The dissimilarity measure higher than a predefined threshold indicates the presence of a shot cut. Experiments have been conducted on the standard TRECVID video database to evaluate the performance of the proposed method.

3.2 VIDEO DATABASE REPRESENTATION TECHNIQUES USING FACES

3.2.1 FACE REPRESENTAION BASED ON SIFT-PCA[2]

In this work, a new technique called discriminative SIFT is devised for efficient and accurate face representation and recognition. Unlike scale invariant feature transform model that uses massive dimension of feature vectors, in the proposed model principal component analysis is employed for compact representation of feature vectors. Contrasting to PCA-SIFT model that employs PCA on normalized gradient patch, we employ PCA on smoothed weighted histograms.

Scale Invariant Feature Transform (SIFT) is an approach for detecting and extracting local feature descriptors that are reasonably invariant to changes in illumination, image noise, rotation, scaling, and small changes in viewpoint. The SIFT transforms image data into scale-invariant coordinates relative to local features that consists of four major phases: (1) scale-space peak selection; (2) keypoint localization; (3) orientation assignment; (4) keypoint descriptor.

The first stage of SIFT model is scale-space extrema detection, which computes interest points that are invariant to scale and orientation are extracted by searching over all scales and image locations. Interest points for SIFT features correspond to local extrema of difference-of-Gaussian filters at different scales. The first step toward the detection of interest points is the convolution of the image with Gaussian filters at different scales, and the generation of difference-of- Gaussian images from the difference of adjacent blurred images. The convolved images are grouped by

[2] *Discriminative SIFT model: A robust and an accurate face representation and recognition model,* ***International Journal of Information Processing,*** *vol. 4(2), pp. 51-61, 2010.*

Discriminative SIFT model: A robust and an accurate face representation and recognition model, ***International Conference on Information Processing 2009 (ICIP-09),*** *pp. 178-190, UVCE, Bangalore, India. 6-9 Aug. 2009.*

SIFT-PCA: An efficient and accurate face representation and recognition model, ***International Conference on Artificial Intelligence, IICAI-2009****, pp. 1969-1980, Dec. 16-18, 2009.*

octave by having same number of difference-of-Gaussian images per octave with a fixed number of blurred images per octave.

The second stage is keypoint localization, where keypoints are identified as local maxima or minima of the DoG images across scales. Each pixel in the DoG images is compared to its 8 neighbors at the same scale, plus the 9 corresponding neighbors at neighboring scales. If the pixel is a local maximum or minimum, it is selected as a candidate keypoint. All the candidate keypoints are localized to sub-pixel accuracy and eliminated if found unstable.

The third stage is orientation assignment, where the keypoint is assigned an orientation. To determine the keypoint orientation, a gradient orientation histogram is computed in the neighborhood of the keypoint using the Gaussian image at the closest scale to the keypoint's scale. The contribution of each neighboring pixel is weighted by the gradient magnitude and a Gaussian window with a σ that is 1.5 times the scale of the keypoint. Peaks in the histogram correspond to dominant orientations. A separate keypoint is created for the direction corresponding to the histogram maximum. All the properties of the keypoint are measured relative to the keypoint orientation which provides invariance to rotation.

Finally, keypoint descriptors are computed by having keypoint orientation. The feature descriptor is computed as a set of orientation histograms on 4 x 4 pixel neighborhoods. The orientation histograms are relative to the keypoint orientation and the orientation data comes from the Gaussian image closest in scale to the keypoint's scale. Histograms contain 8 bins each, and each descriptor contains an array of 4 histograms around the keypoint. This leads to a SIFT feature vector size (4x4) **x** 8. Although the standard SIFT descriptor is significant in representing the image through keypoints, the dimension of the feature vector is quite high and hence time consuming in real-world applications. Hence, we have proposed an alternate representation scheme that uses PCA for dimensionality reduction and thus extracts only meaningful and discriminative information from the keypoints and provides a succinct representation associated with each keypoint.

The proposed model posses invariance to linear transformation and robust to occlusion and noisy environment. The proposed model has better computing performance in terms of recognition time than the basic SIFT technique. To establish the superiority of the proposed model, we have experimentally compared the performance of our new algorithm with recently proposed $(2D)^2$-PCA on the benchmark databases: AT&T and UMIST face databases.

3.2.2 FLD-SIFT: CLASS BASED SCALE INVARIANT FEATURE TRANSFORM FOR ACCURATE CLASSIFICATION OF FACES [3]

In this work, a new model called FLD-SIFT is devised for compact representation and accurate recognition of faces. Unlike scale invariant feature transform model that uses smoothed weighted histogram and massive dimension of feature vectors, in the proposed model, an image patch centered around the keypoint has been considered and linear discriminant analysis (FLD) is employed for compact representation of image patches. Contrasting to PCA-SIFT model that employs principal component analysis (PCA) on a normalized gradient patch, we employ FLD on an image patch exists around the keypoints.

The proposed model has better computing performance in terms of recognition time than the basic SIFT model. To establish the superiority of the proposed model, we have experimentally compared the performance of our new algorithm with $(2D)^2$-PCA, $(2D)^2$-FLD and basic SIFT model on the AT&T face database.

[3] *FLD-SIFT: Class Based Scale Invariant Feature Transform for Accurate Classification of Faces,* ***Second International Conference on Advances in Communication, Network, and Computing – CNC-2011****, V.V. Das, J. Stephen & Y. Chamba (Eds.), CCIS-142, Springer-Verlag, pp. 15-21, Mar. 2011.*

3.2.3 FLD-SIFT: CLASS BASED SCALE INVARIANT FEATURE TRANSFORM FOR ACCURATE CLASSIFICATION OF FACES[4]

It is observed from the literature that the monogenic signal requires fewer convolutions when compared to Gabor wavelet based model and hence require much lesser computing time. The region covariance matrix descriptor is employed on monogenic scale space based images to obtain the representation of face images. The monogenic signal is based on the Riesz transform which is used instead of the Hilbert transform. The monogenic signal analysis is a framework to interpret images in terms of the local phase, local orientation and local energy. The monogenic signal is an effective tool to analyze 2-D signals in a rotation invariant manner. The signal is built upon the first order Riesz transform. Hence, the energy and phase information contained in the monogenic signal at different scales is used to build the region covariance matrix descriptor for each face image. Experimental results on the standard benchmark databases reveal the superiority of the proposed model for face recognition problems and its suitability in real environment.

[4] *Monogenic Scale Space Based Region Covariance Matrix Descriptor for Face Recognition, DST-FRBR sponsored Russian-Indian Bilateral workshop on 'Emerging Applications of Computer Vision', pp. 29-36, Nov. 2011.*

3.2.3 FACE REPRESENTATION USING KERNEL ENTROPY COMPONENT ANALYSIS[5]

We have developed a new face recognition algorithm based on Renyi entropy component analysis (Jensen, 2010). The proposed model integrates kernel-based methodology with entropy analysis to choose the best principal component vectors which are subsequently used for pattern projection to a lower-dimensional space.

The kernel PCA (Schokopf et al., 1999) is perhaps the best known data transformation method in a non-linear sense. Let $\mathbf{X} = [x_1, x_2, \ldots, x_N]$, where $x_i \in R^d$, i=1,…,N. The basic idea of kernel PCA is to map the input data **X** into a feature space F via a nonlinear mapping ϕ and then perform a linear PCA in F.

The estimated covariance matrix of the mapped data $\phi(x_i)$ in kernel PCA is defined as:

$$C = \frac{1}{N}\sum_{i=1}^{N} \Phi(x_i).\Phi(x_i)' \qquad \ldots (1)$$

And the corresponding eigenvalue problem is:

$$\lambda w = Cw.$$

In the case of kernel PCA, the top α_l eigenvectors associated with first l large eigenvalues are considered to transform high dimensional face image data into its lower dimensional data. Although kernel ECA is similar to kernel PCA in terms of obtaining the eigenvectors and eigenvalues, the procedure of feature extraction is based on the energy content of the eigenvectors and eigenvalues and the details are brought out below.

---

[5] *Face Recognition Using Kernel Entropy Component Analysis,* ***International Journal of Neurocomputing****, 74(6): 1053-1057 (2011).*

Kernel Entropy Component Analysis Based Model For Face Recognition, ICVGIP, Dec. 2010, IIT Madras.

The Renyi quadratic entropy is given by

$$H(p) = -log \int p^2(x)dx, \quad \ldots (2)$$

where p(x) is the probability density function generating the data set, or sample, S = x_1, x_2, …, x_N. Because of the monotonic nature of logarithmic function, one can consider the quantity:

$$V(p) = \int p^2(x)dx \quad \ldots (3)$$

The estimation of V(p) will be done using the Parzen window density estimator as suggested in (Parzen, 1962) and is given below.

$$\hat{p}(x) = \frac{1}{N} \sum_{x_t \in S} k_\sigma (x, x_t) \quad \ldots (4)$$

Here, $k_\sigma(x, x_t)$ is the kernel centered at x_t and width governed by the parameter σ.

Hence,

$$\hat{V}(p) = \frac{1}{N} \sum_{x_t \in S} \hat{p}(x_t)$$

$$= \frac{1}{N} \sum_{x_t \in S} \frac{1}{N} \sum_{x_t \in S} k_\sigma (x_t, x_{t'}) \quad \ldots (5)$$

$$= \frac{1}{N^2} \mathbf{1}^T \mathbf{K} \mathbf{1}$$

Here, the element (t, t') of the N x N kernel matrix K is $k_\sigma(x_t, x_{t'})$ and **1** is an (N x 1) vector containing all ones.

Hence, the Renyi entropy estimator may be expressed in terms of the eigenvalues and eigenvectors of the kernel matrix, which may be decomposed as $\mathbf{K} = \mathbf{E}\mathbf{D}\mathbf{E}^T$ where **D** is the diagonal matrix containing the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_N$ and **E** is a matrix with the corresponding eigenvectors $\alpha_1, \alpha_2, \ldots, \alpha_N$ as columns.

Rewriting the above, we have:

$$\hat{V}(p) = \frac{1}{N^2}\sum_{i=1}^{N}(\sqrt{\lambda_i}\alpha_i^T\mathbf{1})^2 \qquad \ldots (6)$$

The Eq.(6) reveals that the Renyi entropy estimator is composed of projections onto all the kernel PCA axes, however, only a principal axis α_i for which $\lambda_i \neq 0$ and $\alpha_i^T\mathbf{1} \neq 0$ contributes to the entropy estimate. Hence, kernel ECA is defined as an *l*-dimensional data transformation technique obtained by projecting $\boldsymbol{\phi}$ onto a subspace E_l spanned by those *l* kernel PCA axes contributing most to the Renyi entropy estimate of the data. Hence E_l is composed of subset of kernel PCA axes but not necessarily those corresponding to the top *l* eigenvalues.

We have conducted experimentation using YALE and UMIST face image databases. The classification stage is implemented by feeding feature vectors into a nearest neighbour rule and the average error rate, say E_{ave}, is calculated as follows.

$$E_{ave} = \frac{\left(\sum_{i=1}^{r} t_{mis}^{i}\right)}{(r.t)} \qquad \ldots (7)$$

where *r* is the number of runs, *t* is the number of total test samples of each run and t_{mis}^{i} is the number of misclassification for the i^{th} run. It is found from the experimentation that the proposed model has lesser error rate when compared to kernel PCA based face recognition technique.

Extensive experimentation on Yale and UMIST face database has been conducted to reveal the performance of the entropy based principal component analysis method and comparative analysis is made with the kernel principal component analysis method to signify the importance of selection of principal component vectors based on entropy information rather based only on magnitude of eigenvalues.

3.2.4 KERNEL DISCRIMINATIVE EMBEDDING TECHNIQUE WITH ENTROPY COMPONENT ANALYSIS FOR ACCURATE FACE IMAGE CLASSIFICATION [6]

In this work, we have reported a new face image classification algorithm based on Renyi entropy component analysis. In the proposed model, kernel discriminant analysis is integrated with entropy analysis to choose the best principal component vectors which are subsequently used for pattern projection to a lower dimensional space.

The kernel FLD (Lu et al., 2003) is perhaps the best known data transformation method in a non-linear sense when class-wise information is available. Let $X = [x_1, x_2, \ldots, x_N]$, where $x_i \in R^d$, i=1,…,N represent the set of N face images converted into a vector form of *d* dimension. It is further assumed that each image belongs to one of C classes. The basic idea of kernel FLD is to map the input data X into a feature space F via a nonlinear mapping ϕ which is given by $\phi: R^d \rightarrow F$ such that $x_t \rightarrow \phi(x_t)$, t=1,…,N and then to perform an LDA in the feature space F instead of input space R^d. Let $\phi = [\phi(x_1), \phi(x_2), \ldots, \phi(x_N)]$.

Let S_B and S_w be the between-class and within-class scatter matrices in the feature space F, respectively, expressed as follows:

$$S_B = \frac{1}{L}\sum_{i=1}^{C} C_i\left(\overline{\phi_1} - \emptyset_i\right)\left(\overline{\emptyset_1} - \emptyset_i\right)^T$$

$$S_W = \frac{1}{L}\sum_{i=1}^{C}\sum_{j=1}^{C_i} \left(\emptyset_{ij} - \overline{\emptyset_1}\right)\left(\emptyset_{ij} - \overline{\emptyset_1}\right)^T$$

Where $\emptyset_{ij} = \emptyset(x_{ij})$, $\overline{\phi_1} = (1/C_i)\sum_{j=1}^{C_i}\emptyset(x_{ij})$, is the mean of the class C_i, $\overline{\phi_1} = \frac{1}{L}\sum_{i=1}^{C}\sum_{j=1}^{C_i}\emptyset(x_{ij})$ is the average of the ensemble, and X_i is the element number in Z_i, which leads to $L = \sum_{i=1}^{C}\sum_{j=1}^{C_i} C_i$. The LDA determines a set of optimal

[6] *Kernel Discriminative Embedding Technique with Entropy Component Analysis for Accurate Face Image Classification*, ***International Conference on Image Processing***, ***ICIP-2011,*** *K R Venugopal and L M Patnaik (Eds.), CCIS-157, Springer-Verlag, pp. 181-189*, *Aug. 2011.*

discriminant basis vectors, denoted by $\{\psi_k\}_{k=1}^{N}$, so that the ratio of the between class scatter matrix and within class scatter matrix is maximized (Lu et al., 2003). As summing $\psi = [\psi_1,\ldots,\psi_N]$,the maximization can be achieved by solving the following eigenvalue problem:

$$\psi = argmax \frac{|(\psi^T S_B \psi)|}{|(\psi^T S_W \psi)|}$$

Subsequently, we project the vectors in R^d to a lower dimensional space spanned by the eigenvectors ψ. Let x be a test sample whose projection is $\phi(x)$ in R^d, then the projection of $\phi(x)$ onto the eigenvectors ψ is the non-linear Fisher's principal components corresponding to ϕ:

$$\psi.\phi(x) = \sum_{i=1}^{m} \alpha_i\big(\phi(x_i).\phi(x)\big) = \sum_{i=1}^{m} \alpha_i k(x_i, x)$$

In other words, we can extract the first q non-linear principal components which are corresponding to the first q non-increasing eigenvalues using the kernel function.

In the case of kernel FLD, the top α_q eigenvectors associated with first q large eigenvalues are considered to transform high dimensional face image data into its lower dimensional data. Although kernel ECA is similar to kernel FLD in terms of obtaining the eigenvectors and eigenvalues, the procedure of feature extraction is based on the energy content of the eigenvectors and eigenvalues.

Extensive experimentation on Yale and UMIST face database has been conducted to reveal the performance of the entropy based kernel discriminative embedding technique and comparative analysis is made with conventional kernel linear discriminant method to signify the importance of selection of principal component vectors based on entropy information rather based only on magnitude of eigenvalues.

3.3 VIDEO DATABASE REPRESENTATION TECHNIQUES USING TEXT DATA

In the current decade, we have seen the rapid growth of multimedia based applications and hence there is a need for the design and development of efficient retrieval systems which process the massive amount of multimedia (video, audio, text) data. In this context, much effort has been done by the researchers to use either image or text as a means for indexing large multimedia data for quicker retrieval. It is observed that the detection of texts in video without any constraints remain challenging and interesting due to many undesirable properties of video frames such as low resolution, low contrast, unknown text colour, size, position, orientation, colour bleeding and unconstrained background (Jung et al., 2004; Ye et al., 2005). In general, we have seen two categories of texts in video frames: c*aption/graphics/artificial* text which is artificially superimposed on the video by human and *scene* text which naturally occurs during video capture. Noticeably, scene text detection is a challenging task compared to graphics text due to varying lighting conditions, complex background and geometrical transformations (Zang and Kasturi, 2008).

Several models have been proposed by many researchers to detect text from video frames which are broadly classified as bottom-up models (Wernicke and Lienhart, 2002; Clark and M Mirmehdi, 2000); top-down models (Sobottka and Bunke, 1999; Hanif and Prevost, 2009) and machine learning based models (Jain et al., 1995; Lienhart, 1996). In the bottom-up based approaches, probable character regions are identified in a frame and then groups character regions into words by using geometrical constraints such as the size of the region, height and width ratio. In the second category, character attributes such as edge orientation, texture values, spatial information is used to localize texts. The machine learning based models use statistical measures (Li and Doermann, 2000; Clark and Mirmehdi, 2000), multi-scale analysis (Jain et al., 1995) of video frame to detect and localize the textual information in video frames and argued that these categories of models exhibit better performance.

In our work, we have proposed the following text detection models which are subsequently used for indexing of video documents.

3.3.1 TEXT LOCALIZATION BASED ON WEBER'S LOCAL DESCRIPTOR AND CLUSTERING [7]

In this paper, we have proposed an accurate text detection model based on Weber's local descriptor (WLD) introduced by Chen (2009). WLD is based on the fact that human perception of a pattern depends on not only the change of a stimulus but also the original intensity of the stimulus. More specifically, WLD consists of two components: its differential excitation and orientation. A differential excitation is defined as the ratio between relative intensity differences of the current pixel (pixel under processing) and the intensity of the current pixel. An orientation is the gradient orientation of the current pixel. Given a document image, we compute the differential excitation and orientation value of each pixel and their difference is computed to highlight the text components in a document image. Subsequently, K-means clustering based connected component analysis is performed on the processed document image to detect text from a document image. The proposed scheme is evaluated on documents taken from ICDAR 2003 robust reading and text locating database. We have made a comparison with the other existing contemporary models to exhibit the performance of the proposed model.

[7] *Weber's local descriptor based model for text detection in document images,* ***International Conference on Information Processing-2010 (ICIP-2010),*** *Bangalore, India. Aug. 06-08, pp. 426-436, 2010.*

Weber's local descriptor based model for text detection in document images, ***International Journal of Information Processing,*** *Vol. 4(4), pp. 17-25, 2010.*

3.3.2 TEXT LOCALIZATION BASED ON WEBER'S LOCAL DESCRIPTOR AND DIFFERENTIAL EXCITATION DIFFERENCE [8]

In this work, we have proposed a robust and an efficient text detector in a video frame. The proposed model is based on Weber's excitation principle and difference of excitation information is used to highlight the textual data in a video frame. In order to exhibit the performance of the proposed model, we have conducted experimentation on the standard ICDAR-2003 dataset and our own video database. Experimental comparison and objective comparison is also provided with the recently proposed models.

[8] *Text Detection in Video Frames: An Integrated Approach Based On Weber's Local Descriptor And Differential Excitation Difference,* ***MES Journal of Technology and Management****, Vol. no. 1(2), pp. 76-82, 2010.*

Text Detection in Video Frames: An Integrated Approach Based On Weber's Local Descriptor And Differential Excitation Difference, ***International Conference on Computational Engineering, Practices and Techniques,*** *pp. 76-82, Nov. 25-26, 2010, Kerala.*

3.3.3 TEXT LOCALIZATION BASED ON MORAVEC OPERATOR[9]

In this work, we have proposed an accurate and an efficient text detection model for document images and video frames. The proposed model is based on the Moravec operator to localize text in a document image and horizontal and vertical mask convolution is performed to identify the actual text blocks. In order to exhibit the performance of the proposed model, we have conducted experimentation on the standard ICDAR-2003 dataset and also on the video frames collected from internet. The text blocks of video frames possess properties such as varying fonts, colour, size, languages etc. The superiority of the proposed text detection model is demonstrated through experimental comparison and objective comparison with the recently proposed text detection algorithms.

[9] *On the use of Moravec Operator for Text Localization in Web Documents,* ***IEEE International Conference on Recent Trends in Information Technology****, pp. 910-914, June 2011.*

3.3.4 TEXT DETECTION IN DOCUMENT IMAGES AND VIDEO FRAMES BASED ON PHASE OF THE FOURIER TRANSFORM[10]

In this paper, we have proposed an accurate and an efficient text detection model for document images and video frames. The proposed model is based on the phase of the 2-D Fourier transform of the document image/video frame to overcome the adverse effect of illumination. The proposed model better localize the text blocks present in a document image as the phase of the Fourier transform preserves the locations of the edges of a given document image and suitable post-processing mechanism is devised to correctly identify the text blocks. However, the main problem in the use of the phase spectrum is the need for unwrapping of the phase. And hence the problem of unwrapping is avoided by considering two functions of the phase spectrum rather than the phase directly. Each of these functions gives partial evidence of the presence of text in a document image. In order to exhibit the performance of the proposed model, we have conducted experimentation on the standard ICDAR-2003 dataset and also on the video frames collected over the internet. The text blocks of video frames possess properties such as varying illumination, varying fonts, colour, size, languages etc. Extensive experimental comparison and objective comparison is provided to exhibit the performance of the proposed text detection model.

[10] *Text Detection in Document Images and Video Frames: An Accurate and Robust Approach Based on Phase of the Fourier Transform,* ***International Conference on Frontiers of Computer Science****, Aug. 2011.* ***[BEST PAPER AWARD]***

3.3.5 PHASE OF THE DISCRETE COSINE TRANSFORM FOR TEXT DETECTION IN DOCUMENT IMAGES AND VIDEO FRAMES [11]

In this work, we have proposed an illumination invariant and an efficient text detection model for document images and video. The proposed model is based on the phase of the 2-D discrete cosine transform of the document image/video frame to overcome the adverse effect of illumination. The proposed model better localize the text blocks present in a document image as the phase of the cosine transform preserves the locations of the edges of a given document image and suitable post-processing mechanism is devised to correctly identify the text blocks. In order to exhibit the performance of the proposed model, we have conducted experimentation on the standard ICDAR-2003 dataset and also on the video frames collected over the internet. The text blocks of video frames possess properties such as varying illumination, fonts, colour, size, languages etc. Extensive experimental comparison and objective comparison is provided to exhibit the performance of the proposed text detection model.

[11] *On the use of phase of the discrete cosine transform for text detection in document images and video frames under variations in illumination,* ***Indian International Conference on Artificial Intelligence****, Dec. IICAI 2011.*

3.3.6 INDEXING VIDEO DATABASE USING UNIVERSAL B-TREE [12]

The developments in the digital and network technology have produced a downpour of information. The amount of video information in particular has led to unprecedented high volumes of data. When fast-forwarding through videotape, a user searches for an image or sequence similar to that in their imagination. In some complex cases queries are not this simple, but a system that can locate and present keys relevant to the video content-instead of depending on the user's imagination-will promote easier handling of extensive videos. The essential issues involve assisting users by extracting physical features from video data and designing an effective application interface. We can extract physical features to partition video data into useful segments (shots) and store the segment attribute information, or annotations, as indexes. These indexes should describe the essential information pertaining to the video segments, and should be content based. Indexes can be visualized through the interface so users can perform various functions. Generally, these indexes are of multi-dimensional and hence require a development of appropriate access methods. Unlike single-dimensional databases, which are indexed/searched according to a simple key (e.g. using B-trees), we often intend to access data according to a composite key (according to several attributes generally). We call such databases multi-dimensional, since data instance is represented by a vector of simple values. A collection of data vectors (data tuples) can be interpreted as set of points in a multi-dimensional vector space.

With this backdrop, we have made an attempt to explore the suitability of universal-B tree (Bayer, 1997; Markl et al., 2000) a multi-dimensional indexing tree structure, to store key frames and hence the desired information can be quickly retrieved with the help of range queries by which the user specifies an interval of values $<a_i, b_i>$ (for each attribute A_i), which the retrieved data tuples have to match.

The proposed indexing has better computing performance in terms of retrieval time than the basic sequential/other indexing techniques.

---

[12] *Under preparation - Universal B-Tree based video indexing for accurate retrieval of video frames,* ***IEEE Transactions on Knowledge and Data Engineering.***

CONCLUSION

In this research work, we have devised some novel methodologies for video database analysis, indexing and retrieval. The research work is focused on devising algorithmic models which are efficient and accurate to represent video database. The proposed models are broadly classified based on the *textual information* of video document and *human faces* information present in a video frame.

In the first part, representation techniques based on visual features (faces) are developed. The methods presented are based on a structural decomposition strategy where the video database is segmented into shots which are subsequently used to identify faces in a key frame. We have devised novel and efficient representation schemes based on subspace analysis techniques. In addition, we have also devised techniques based upon local descriptors useful in many computer vision applications. Many of the problems with the existing subspace analysis models are eliminated in the proposed models. Higher attention is given to reduce the computational burden, memory requirements and finally towards providing an added degree of freedom. Extensive experiments have been conducted on AT&T, UMIST and YALE face databases to exhibit the performance of the proposed representation schemes.

In the second part of our work, we have addressed the problem of video frame representation using textual information. We have devised many text localization models which are capable of localizing the text accurately and hence useful for representation of the video frames using text data.

The multi-dimensional indexing data structure called Universal B-Tree is introduced for the purpose of storing the video contents and hence analysis is provided to justify the usage of indexing mechanism for content based video retrieval applications.

REFERENCES

Bayer, R, 1997. The universal B-tree for multidimensional indexing: General Concepts. World Wide Computing and its Applications, Tsukuba, pp. 10-11.

Chen, J., Shiguang Shan, Chu He, Guoying Zhao, Matti Pietikainen, Xilin Chen, Wen Gao. WLD: A Robust Local Image Descriptor, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 06 Aug. 2009.

Clark, P. and M Mirmehdi. Combining statistical measures to find image text regions. *In Proceedings of the 15th International Conference on Pattern Recognition*, pp. 450-453. IEEE Computer Society, September 2000.

Clark, P. and M Mirmehdi. Finding text regions using localized measures. *In Proceedings of the 11th British Machine Vision Conference*, pages 675-684. BMVA Press, September 2000.

Hanif, S. M. and L Prevost. Text detection and localization in complex scene images using constrained AdaBoost algorithm, *10th International Conference on Document Analysis and Recognition*, pp. 1-5, 2009.

Jain, A.K., Y. Zhong and K. Karu, 1995. Locating text in complex color image. Pattern Recognition, pp. 1523-1536.

Jenssen, R., *Kernel entropy component analysis*, IEEE Transactions on PAMI, vol. 32(5), pp. 847-860, 2010.

Jung, K., K.I. Kim and A.K. Jain. "Text information extraction in images and video: a survey". *Pattern Recognition,* 37, 2004, pp. 977-997.

Li, H. and D Doermann. A video text detection system based on automated training. *In Proceedings of the International Conference on Pattern Recognition*, ICPR'00, 2000.

Li, Y. and C. J. Kuo, Video Content Analysis Using Multimodal Information: For Movie Content Extraction Indexing and Representation, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2003.

Lienhart, R., Automatic text recognition in digital videos. *In SPIE, Image and Video Processing IV*, pp. 2666-2675, 1996.

Lienhart, R., Comparison of automatic shot boundary detection algorithms, *Proc. of SPIE*, vol. 3656, pp. 290-301, January 1999.

Lu, J., K.N. Plataniotis, and A. N. Venetsanopoulos. *Face recognition using kernel direct discriminant analysis algorithms.* IEEE Transactions on Neural Networks, vol. 14 (1), pp. 117-126, 2003.

Manjunath, B., P. Salembier, T. Sikora, and P. Salembier, 2002. Introduction to MPEG 7: Multimedia Content Description Language. John Wiley & Sons.

Parzen, E., *On the estimation of a probability density function and the mode*. The Annals of Math. Statistics, vol. 32, pp. 1065-1076, 1962.

Rui, Y., T. S. Huang, and S. Mehrotra, "Constructing table-of-content for videos," Multimedia Systems, vol. 7, no. 5, pp. 359–368, 1999.

Scholkopf, B.S., Mika, C.J.C Burges, P. Knirsch, K.-R Muller, G. Ratsch, and A.J. Mola, 1999. Input space versus feature space in kernel based methods, IEEE Transactions on Neural Networks, vol. 10(5), pp. 1299-1319.

Sobottka, K. and H Bunke. Identification of text on colored book and journal covers. *In International Conference on Document Analysis and Recognition*, pages 57-62, Bangalore, India, September 1999.

Tsekeridou, S. and I. Pitas, Content-based video parsing and indexing based on audio-visual interaction, *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 4, pp. 522{535, 2001.

Wernicke, A. and R Lienhart, 2002. Localizing and segmenting text in images and videos. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 18(8), pp. 256-268.

Ye, Q., Q. Huang, W. Gao and D. Zhao. "Fast and robust text detection in images and video frames". *Image and Vision Computing* 23, 2005, pp. 565-576.

Yeung, M.M. and B. L. Yeo, Video content characterization and compaction for digital library applications, *Proc. of SPIE*, vol. 3022, pp. 45{58, February 1997.

Zhang, H. J., C. Y. Low, S. W. Smoliar, and J. H. Wu, Video parsing, retrieval and browsing: An integrated and content-based solution, *ACM Multimedia'95*, November 1995.

Zhang, J. and R. Kasturi. "Extraction of Text Objects in Video Documents: Recent Progress". *The Eighth IAPR Workshop on Document Analysis Systems (DAS2008)*, Nara, Japan, September 2008, pp 5-17.

Zhang, T. and C.-C. J. Kuo, Audio-guided audiovisual data segmentation, indexing and retrieval, *Proc. of SPIE*, vol. 3656, pp. 316-327, 1999.

SHARMILA KUMARI, M. — Research Scholar

Dr. B H SHEKAR — Research Supervisor